09/787323
Rec'd PCT/PTO 16 MAR 2001
PCT/US99/21092

Multiplex Sequencing Method

Background

Field of the Invention

The present invention provides a method for identifying a nucleic acid utilizing a run-off sequencing reaction of a relatively short portion of the nucleic acid. The method can be utilized, for example, to identify an EST from only a small portion of the EST and in an analysis of nucleotide polymorphisms. The reactions can be multiplexed to increase data readout capacity.

Background of the Invention

Several methods have been developed to increase the efficiency of DNA sequencing analysis. These include the methods of i) multiplexing a series of spectrally non-overlapping terminator and/or dye-primer dyes into DNA sequencing lanes, ii) transfer of genomic sequencing reactions to a filter and subsequent hybridization, and iii) multiplex lane-loadings in which 3 instead of 4 sequencing reactions are performed. These methods have mainly been applied to situations in which a long read (greater than several hundred bases *de novo*) is desired.

The present invention is the development of a simple method for multiplexing short sequencing reads (about 16 bases) in the same lane. The application to which we are applying this method is our high-throughput yeast two-hybrid analysis (Buckholz, Stuart, Judelson and Weiner). In this analysis, we desire to sequence short regions of the interacting proteins, and then use a large database to determine the hit identification. Because each bait analyzed generates approximately 100 hits, we needed to develop a method to increase our efficiency of analysis.

Description of the Figures

ins. B2

Figure 1. Untreated and *Bpm*I-treated sequencing reaction. See text for details.

Figure 2. Separation as a function of delta loading time. *Bpm*I-treated PCR fragments were sequenced and multiplexed on the ABI 377 at loadings 1, 2 and 3 at the times indicated post first-loading.

Figure 3. Multiplex loading of a sequencing gel and chromatogram of a single multiplexed lane. Note the chromatogram is not from a lane on the gel shown.

Detailed Description of the Invention

We have developed a method whereby we use reloading of a nucleotide base-calling apparatus, for example polyacrylamide gel electrophoresis or capillary electrophoresis to serial multiplex DNA base-calling. In one example, a run-off sequencing reaction is used to sequence the bases downstream from an endonuclease recognition site. In this method, the endonuclease selected is one that cuts several bases downstream of its recognition site, such that nucleotides from outside the recognition sites would be included in the restricted section of DNA and would then be sequenced in a short-run, run-off sequencing reaction. A short sequencing reaction can be one of 30 or fewer bases, such as 30 bases, 25 bases, 20 bases, 19 bases, 18 bases, 17 bases, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, or 5 bases. For example, in a specific example we term '*BpmI* sequencing,' a run-off sequencing reaction is performed to sequence the 16 bases downstream from a type IIS endonuclease *Bpm*I recognition site.

For this method, a library is constructed wherein the inserts of the library are positioned within the library vector in sufficient proximity to a selected

enzyme recognition site, of an enzyme that cuts downstream of its recognition site, that the insert will be cut by the enzyme within the insert. For example, a library can be constructed from inserts having ligated to them linkers providing the recognition site for the selected enzyme. By way of another example, the vector in which the library is constructed can contain within its multiple cloning site a recognition site for the selected endonuclease to be used to create the template for the run-off sequencing reaction, and the library inserts can be cloned into the vector in a site such that the inserts are in sufficient proximity to the recognition site of the selected enzyme such that the inserts will be cut by the selected enzyme. Furthermore, primers can also be designed to allow amplification of an isolated subclone of the library, prior to performing the restriction and sequencing reactions, wherein the restriction recognition site of the selected endonuclease is retained within the amplified region.

One advantage to this invention is that sequencing reactions can be multiplexed on the analysis apparatus, because it produces short sequences. The sequences determined, however, are sufficient for identifying the isolated nucleic acid by comparison with a sequence database. Thus, for example, two, three, four or more sequences can be run sequentially on the analysis apparatus, allowing for a significant decrease in time and cost of obtaining the data.

One utility of this method is in comparing sequenced cDNA against a cDNA database, for example GenBank. Given such a comprehensive cDNA database, it should be possible to determine the identification of an EST from an analysis of just a small portion of the EST. We are applying this technology to yeast two-hybrid (Y2H) analysis of protein-protein interactions in which a known bait-protein fusion is tested for interactions with an expressed cDNA library. To test the *Bpm*I sequencing method, we cloned randomly primed macrophage cDNA into a yeast two-hybrid cDNA library vector using adapters incorporating a *Bpm*I restriction endonuclease recognition site. Clones have been isolated from the library and tested for the correct gene-call after *Bpm*I sequencing. By sequencing just a small region of DNA adjacent to the cloning site, one can multiplex the DNA sequencing reactions and thereby increase the gene readout capacity of most analytical methods.

Another example of a utility of this invention is in the use of multiplexing sequencing runs applied toward SNP analysis whereby short PCR products containing the region-of-interest are loaded repeatedly into the same well/capillary tube and sequentially-analyzed.

The present method for run-off DNA sequencing can be used to increase the sequencing capacity of a single gel several fold. For example, the *Bpm* I method for run-off DNA sequencing can be used to increase the sequencing capacity of a single gel at least 4 fold. A 16 bp read from one end of the clone can be used to correctly identify many clones. With the implementation of Bioinformatics tools such as sample tracking software and a tool to merge the BLAST results of the forward and reverse reactions, this methodology can be used to support Y2H in a higher-throughput environment.

The enzyme utilized to cut the nucleic acid sample for sequencing is an enzyme that cuts at least 1 base downstream of its recognition site, so that the run-off sequencing event produces sequence data including the nucleotide sequences of the library insert up to the point of restriction by the enzyme. Thus, the enzyme can be a restriction endonuclease. In addition to *Bpm* I, exemplified herein, which cuts 16 bases downstream of its recognition site, other non-palindromic endonucleases such as *Bsg* I (16/14) and *Eco*57 I (16/14) can readily be used to design linkers for run-off sequencing. For further example, *Bcg* I, *Fok* I, or another enzyme which would allow a longer read, *Mme* I (20/18), could be utilized. The enzyme can be chosen by considering the number of bases of sequence data desired for the specific purpose.

Additional optimization of this technology can be done. Redesign of sequencing primers to read closer to the cloning site will allow for shorter sequencing reads and increase the multiplexing capacity of the gel. Additionally, longer run times on the ABI 377XL may have an advantage. Furthermore, a system featuring automated sample loading, such as the ABI 310 can be utilized.

Analysis may be performed by any means desired. For example, analysis of gel electrophoresis, analysis on a capillary apparatus, or analysis by mass spectrophotometry can be performed.

Also provided is a kit for performing multiplex analysis of sequencing reactions comprising: an enzyme that cuts at least 1 base downstream of a selected enzyme recognition site; and a set of oligonucleotide linkers comprising a recognition site for the selected enzyme. For example, the enzyme can be *Bpm I*, *Bsg I*, *Eco57*, or *MmeI* or a combination thereof. The kit can further comprise, for example, a vector for constructing a library wherein, for example, the vector has an appropriate cloning site for use in the method. The kit can further comprise a component to facilitate the multiplexing of the sequence reaction products, selected according to the analysis method to be used.

Examples

cDNA library construction. Polyadenylated RNA was isolated from 5 x 10^7 THP1 cells using FastTrack 2.0 (Invitrogen, San Diego, CA). A random oligomer primed cDNA library was constructed from 5 μg of the polyA-selected mRNA using the Copy Kit (Invitrogen). *E. coli* DNA ligase was removed from the second-strand synthesis reaction to enhance synthesis of products approximately 900 base pairs in length. Next, *Bpm*I linkers (SEQ ID NO. 1 5'-AATTCGGCTCGAG<u>CTGGAG</u>-3' and SEQ ID NO. 2 5'-<u>CTCCAG</u>CTCGAGCCG-3') were added to the ends of the blunt-ended cDNA fragments using T4 DNA ligase. Following the addition of the linkers, the fragments were phosphorylated (T4 DNA kinase) and size selected using a Chromaspin 400 column (Clontech, Palo Alta, CA). The cloning vector pYesTrp2 (Invitrogen) was digested using the restriction endonuclease *Eco*RI at 37 °C. The linearized vector was dephosphorylated with shrimp alkaline phosphatase (SAP, Boerhinger Mannheim) prior to gel purification. cDNA inserts and treated, linearized vector DNAs were ligated into the cloning vector and the ligation product was transformed into

A
A

Electromax DH10B competent cells (Life Technologies Inc., Gaithersburg, MD). Colonies were selected on LB agar plates with ampicillin.

***Bpm*I sequencing.** Plasmid DNAs were isolated using the R.E.A.L prep (Qiagen, Valencia, CA). One μg of plasmid DNA was digested with 2 U of *Bpm*I (New England Biolabs, Beverly, MA) for at least two hours at 37 °C. Reactions were precipitated with sodium acetate and ethanol, pelleted for 30 min at 3K RPM in a Sorvall RC3B centrifuge rotor. The supernatants were decanted and the pellets were washed with 70% ethanol and dried prior to preparation of sequencing reactions. Using standard conditions, 500 ng of digested DNA was cycle-sequenced using 3.2 pM of primer pYesTrpF or pYesTrpR (Invitrogen) and Big Dye Terminators (PE Biosystems, Foster City, CA). Excess primers and nucleotides were removed using a gel filtration cartridge (Edge Biosystems, Gaithersburg, MD). Products were analyzed on either an ABI 377 or ABI 310 automated sequencer under conditions as specified by the manufacturer and subjected to BLAST analysis against the GenBank database (Table 1).

Table 1. Blast Results

	Undigested		Digested (100% match)	
Clone[a]	Gene Call	E-val	5'-16 Bases	3'-16 Bases
1	Rattus norvegicus RNA helicase	6e-32	--	Didn't cut
6	H. sapiens PAC clone DJ0170019 from 7p15-p21	3e-27	--	--
8	H. sapiens Pig8 mRNA	e-127	+	--
14	H. sapiens mRNA for hnRNPcore protein A1	e-111	+	+
15	Rattus norvegicus unc-50 related protein mRNA	1e-54	+	polyA+[b]
16	Human calmodulin-dependent protein phosp.	e-127	+	+
17	H. sapiens DNA sequence from BAC217C2	e-133	+	+
18	H. sapiens mRNA for putative DNAmethyltrans.	e-140	+	polyA+
21	H. sapiens splicing factor Sip1 mRNA	e-138	+	--
22	Human DNA sequence from cosmid N114B2	2.5	--	--
27	H. sapiens chromosome 16 BAC clone	1e-27	+	polyA+
32	H. sapiens PAC clone DJ0777023 from 7p14-p15	1e-94	--	--
35	Human DNA sequence from PAC417G15	1e-08	--	--
36	Human alpha satellite DNA	4e-39	+	polyA+
37	H. sapiens DNA sequence from BAC 747E2	1e-48	--	--
44	H. sapiens homolog of Nedd5 mRNA	7e-84	+	poor qual. seq.
47	Human kidney mRNA for catalase	e-122	+	+
48	Human DNA sequence "sequence in progress"	4e-83	--	--

[a] mitochondrial (16%), polyA+ only (4%) & cloning vector (2%) hits eliminated

[b] end contained only polyA+ sequence